Hoofdstuk 12. Groepen vergelijken met ANOVA

In dit hoofdstuk wordt gekeken naar methoden om gemiddelden van meerdere groepen met elkaar te vergelijken. De methoden die hierbij gebruikt worden kijken naar een kwantitatieve afhankelijke variabele (vb. inkomen) en een categorische verklarende variabele (vb. etniciteit).

De inferentiële methode die hierbij gebruikt wordt, wordt ANOVA genoemd. Dit staat voor *analysis of variance*.

F-toets
De kern van de analyse is een significantietoets die gebruik maakt van de F-distributie om verschillen waar te nemen tussen diverse populatiegemiddelden.

Het aantal te vergelijken groepen wordt weergegeven met *g*. De populatiegemiddelden zijn dan μ_1, μ_2, …, μ_g. De steekproefgemiddelden zijn $\bar{y}1, \bar{y}2$, …, $\bar{y}g.$ ANOVA toetst hiermee de volgende stellingen:

H0: $\mu_1 = \mu_2 = \ldots = \mu_g$
H1: ten minste twee van de populatiegemiddelden zijn ongelijk.

Om te toets te kunnen uitvoeren moet aan drie voorwaarden worden voldaan. Ten eerste moet van elke groep de populatiedistributie van de afhankelijke variabele normaal zijn verdeeld. Ten tweede moeten de standaarddeviaties van de populatiedistributies van alle groepen hetzelfde zijn. Ten derde moeten de steekproeven onafhankelijk en willekeurig zijn.

De toets statistiek vergelijkt de gemiddelden aan de hand van twee maten van variantie voor elke groep. De eerste gebruikt de variabiliteit tussen elk steekproefgemiddelde ($\bar{y}g$) en het algemene gemiddelde ($\bar{y}$). De tweede gebruikt de variabiliteit binnen elke groep; dus de variabiliteit rondom $\bar{y}1$, $\bar{y}2$, …, $\bar{y}g$. Over het algemeen is het zo dat hoe groter de variabiliteit tussen de steekproefgemiddelden en hoe kleiner de variabiliteit binnen de afzonderlijke groepen, des te meer bewijs tegen H0 dat alle gemiddelden gelijk zijn.
Het getal dat kijkt naar variabiliteit tussen elk groepsgemiddelde ($\bar{y}g$) en het algemene gemiddelde ($\bar{y}$) wordt *between-groups estimate* genoemd. Het getal dat kijkt naar variabiliteit binnen de groepen wordt *within-groups estimate* genoemd. Hiermee wordt de ANOVA F-statistiek berekend, namelijk door de *between-groups estimate* te delen door de *within-groups estimate*. Hoe groter de F-waarde, hoe kleiner de P-waarde.

Within-groups estimate
Deze wordt berekend door de variabiliteit van elke groep te delen door zijn vrijheidsgraden. De formule hiervoor is: $\frac{Within\text{−}groups\ Sum\ of\ Squares}{df}$. De vrijheidsgraden (df) is hier de totale steekproefgrootte minus het aantal groepen. Dus N – g. De *within groups Sum of Squares* kan worden berekend met de volgende formule: $(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2 + \ldots + (n_g - 1)s^2_g$.

Between-groups estimate
Deze kijkt naar de variabiliteit van alle steekproeven ten opzichte van het algemene gemiddelde. De formule hiervoor is: $\frac{Between-groups\ Sum\ of\ Squares}{df}$. De vrijheidsgraden zijn hier het aantal groepen (g) minus 1. Dus g – 1. De *between groups Sum of Squares* kan worden berekend met de volgende formule: $n_1(\bar{y}_1 - \bar{y})^2 + \ldots + n_g(\bar{y}_g - \bar{y})^2$.

Het vergelijken van gemiddeldes
Wanneer bij ANOVA een kleine p-waarde wordt gevonden, geeft deze nog niet aan welke gemiddeldes er verschillen of in welke mate ze verschillen. Daar kunnen betrouwbaarheidsintervallen voor gebruikt worden. Het is informatiever om goed de populatiegemiddelden te schatten, dan om alleen te kijken of ze mogelijk van elkaar verschillen. Er kan voor elk afzonderlijk gemiddelde een betrouwbaarheidsinterval gemaakt worden, maar ook voor het verschil tussen twee gemiddelden. Een betrouwbaarheidsinterval voor het schatten van een populatiegemiddelde kan gemaakt worden met: $\bar{y}_i \pm t\frac{s}{\sqrt{n_i}}$. Een betrouwbaarheidsinterval voor het schatten van het verschil tussen twee populatiegemiddeldes kan gemaakt worden met: $(\bar{y}_i - \bar{y}) \pm ts\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$. Er is een bewezen verschil tussen de twee gemiddeldes wanneer de betrouwbaarheidsinterval geen 0 bevat.

Bovenstaande heeft niet zo veel zin wanneer er heel veel groepen zijn. Dan wordt namelijk de kans ook groter dat er sowieso een resultaat gevonden wordt en er onjuiste inferenties gemaakt gaan worden.

(One-way) ANOVA in een regressiemodel
Multipele regressie kijkt naar de relatie tussen een kwantitatieve afhankelijke variabele en een set van kwantitatieve verklarende variabelen. ANOVA gebruikt in plaats daarvan juist een categorische verklarende variabele, waarvan de categorieën de te vergelijken groepen zijn. Het is te illustreren aan de hand van een regressiemodel.

Bijvoorbeeld er zijn twee mogelijke categorieën (0 en 1) bij twee variabelen (z_1 en z_2). Er ontstaan nu drie groepen: groep een ($z_1 = 1$, $z_2 = 0$), groep twee ($z_1 = 0$, $z_2 = 1$) en groep drie ($z_1 = 0$, $z_2 = 0$). Er kan nu een regressieformule worden gebruikt waarmee het gemiddelde van elke groep kan worden geschat: $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$. Voor elke groep kan deze formule worden ingevuld en het populatiegemiddelde van de groepen worden geschat.

Groep een ($z_1 = 1$, $z_2 = 0$): $E(y) = \alpha + \beta_1$
Groep twee ($z_1 = 0$, $z_2 = 1$): $E(y) = \alpha + \beta_2$
Groep drie ($z_1 = 0$, $z_2 = 0$): $E(y) = \alpha$

Hieruit kan worden opgemaakt dat β_1 het verschil is tussen het geschatte populatiegemiddelde van groep een en groep drie. En dat β_2 het verschil is tussen het geschatte populatiegemiddelde van groep twee en groep drie. Zo wordt duidelijk dat ANOVA een soort regressieanalyse is.

In een ANOVA is de H0-hypothese: µ1 = µ2 = µ3. Als H0 waar is, dan moet µ1 - µ3 = 0, en µ2 - µ3 = 0. Eerder bleek al dat µ1 - µ3 gelijk staat aan β_1 en dat µ2 - µ3 gelijk staat aan β_2. Dus H0 kan ook worden geformuleerd als $\beta_1 = \beta_2 = 0$. De afhankelijke variabele heeft dan voor elke groep dezelfde waarde, namelijk α.

Two-way ANOVA

De one-way ANOVA kijkt naar een kwantitatieve afhankelijke variabele en de categorieën van een enkele verklarende variabele. Met two-way ANOVA kan je categorieën van meerdere verklarende variabelen gebruiken. De nulhypothese is dan dat de populatiegemiddelden hetzelfde zijn over alle categorieën van een van de variabele, terwijl gecontroleerd voor de andere variabelen. Bijvoorbeeld wanneer je wilt weten of mannen en vrouwen evenveel verdienen, terwijl gecontroleerd voor hun politieke voorkeur. De effecten van de individuele predictoren worden ook wel hoofdeffecten (*main effects*) genoemd. Omdat het handmatig uitvoeren van de two-way ANOVA zeer complex is en het doel van het boek te boven gaat, wordt er alleen gebruik gemaakt van computersoftware.

Herhaalde metingen analyse van variantie

De tot nu toe gepresenteerde methoden nemen aan dat de steekproeven in de groepen onafhankelijk zijn: elke groep heeft een aparte steekproef van proefpersonen. In veel studies, echter, hebben de groepen dezelfde proefpersonen. Dit gebeurt vaak wanneer er herhaalde metingen zijn van de proefpersonen over langere tijd of bij meten van gerelateerde responsvariabelen. De steekproeven zijn dan afhankelijk, en dit moet meegenomen worden in de analyse.

One-way ANOVA met Herhaalde Metingen

Neem de tabel op bladzijde 392. Hier is aan 1000 proefpersonen gevraagd wat invloed heeft op de jeugd: Films, programma's op de tv of rockmuziek. Ze konden de invloed aangeven met (erg negatief, negatief, neutraal, positief, zeer positief), dit wordt in de tabel respectievelijk aangegeven met (-2, -1, 0, 1, 2).
H0 is hier: gelijke populatie gemiddelde voor meerdere groepen. Je kunt hier niet normale ANOVA toepassen omdat de 3 steekproeven niet onafhankelijk zijn voor de invloeds-categorieën. Elke steekproef heeft dezelfde proefpersonen.

Stel, we zien de rijen van de tabel als een factor. Dan lijkt de layout van de data op two-way ANOVA.
Een regressiemodel kan de verwachte respons laten zien als een functie van 2 dummyvariabelen voor de 3 invloeden en 11 dummyvariabelen voor de 12 subjecten. De test die populatiegemiddelden voor de 3 invloeden vergelijkt is het hoofdeffect voor de kolomvariabele in een two-way ANOVA.

Uit een tabel van SPSS zijn enkele waarden te halen. De F-test statistic is de mean square voor de invloed gedeeld door de mean square error. Ook kun je de df uit de tabel halen en de P-waarde. Wanneer $P>0.05$, dan is het bewijs tegen H0 niet sterk.

The Sphericity Assumption and Compound Symmetry

De traditionele herhaalde metingen ANOVA, nemen 'sphericity' aan. Dit houdt in dat: voor elk paar groepen, moeten de verschillen tussen twee observaties bekkeen worden, 1 van elke groep. Dit verschil is een variabele, en de sphericity conditie zegt dat de standaard deviatie van de verdeling van dit verschil, identiek is voor elk paar groepen. Het is makkelijker om een bepaalde vorm van sphericity te begrijpen: compound symmetry. In deze conditie hebben de verschillende groepen dezelfde standaarddeviatie, en elk paar van responses hebben dezelfde correlaties. Wanneer deze assumptie niet wordt nagegaan, wordt de P-waarde klein.
Een repeated measurement design kan de precisie van de schatting verbeteren. Het hebben van dezelfde proefpersonen in elke groep helpt om bepaalde bronnen van error tegen te gaan. Bijvoorbeeld: andere variabelen die invloed hebben op de respons hebben dezelfde waardes voor elke groep, dus verschillen tussen groepsgemiddelden duiden niet op verschillen tussen groepen

op deze variabelen. Het controleren van mogelijke confounding factoren, door ze gelijk te houden in elke rij van de data, wordt ‘blocking’ genoemd.

Betrouwbaarheidsintervallen vergelijken afhankelijke steekproeven
In tabel 12.17 werden steekproefgemiddelden gegeven voor de 3 invloeden. De steekproef is klein, dus wordt het multiple comparison betrouwbaarheidsinterval wat zwakker gemaakt, zo dat de intervallen niet te wijd zijn. de 90% Bonferroni BI’s gebruiken kans op error van 0.10/3 = 0.0333 voor elk interval. De error df = 22, en de t-score met kans 0.0333/2=0.0167 in elke staart is t=2.27. De square root van de gemiddelde square error is s = √(0.566) = 0.75. elke groep heeft 12 observaties, dus de errormarge voor elk betrouwbaarheidsinterval is:

Ts √ (+)= 2.27(0.75) √ (+)= 0.70.

Het betrouwbaarheidsinterval voor de verschillen tussen gemiddelde op films en gemiddelde op rock muziek is (-0.08) – (-0.75) ± 0.70, of (-0.03, 1.37). het is mogelijk dat de gemiddelde gelijk zijn, maar ook dat de gemiddelde voor films veel positiever is dan die van rock muziek. In tabel 12.20 zijn de Bonferroni vergelijkingen te zien.

Fixed Effects and Random Effects
Het regressie model voor de vorige analyse is:

$$E(y) = α + β_1 m + β_2 t + y_1 s_1 + y_2 s_2 + … + y_{11} s_{11}.$$

Y is waardering van de invloed, m is de dummyvariabele voor films (bv m = 1 voor respons op films, anders 0), t is een dummyvariabele voor tv (t=1 voor respons op tv, anders 0), en m = t = 0 voor een respons op rock muziek. S_1 is een dummyvariabele voor proefpersoon 1. Y(gamma) wordt in plaats van β gebruikt voor de coëfficiënten, zodat de index van de parameter overeenkomt met de de index van de dummyvariabele. Elke factor heeft 1 dummyvariabele minder dan zijn aantal categorieën.

Een makkelijke manier om dit regressie model op te schrijven is:

$$E(y) = α + β_j + y_i.$$

$β_j$ is hier het effect van invloed j en y_i is het effect voor proefpersoon i, waar $β_3 = 0$ en $y_{12} = 0$ voor de laatste categorie van elke variabele zijn. deze vergelijking laat de verwachte respons zien in de cel in rij i en kolom j, in row main effect en column main effect.

In dit model worden de invloed parameters ($β_j$) vergeleken, en niet de proefpersoon parameters (y_i). De y_i hangt af van welke proefpersonen er zijn gekozen voor de steekproef. Het proefpersoon effect wordt het ‘random effect’ genoemd, omdat de categorieën van de proefpersoon factor een random steekproef representeren. De factor die de groepen definieert, invloed type hier, wordt het ‘fixed effect’ genoemd.

Wanneer de classificatie variabelen een mix zijn van random en fixed effecten, zoals in dit voorbeeld, wordt het model een mixed model genoemd.